21th Southern Statistics Conference:
At 國立中正大學
2011/6/24-25

Regression analysis for dependent truncation data

Emura T* & Wang W (2015), Semiparametric inference for an accelerated failure time model with dependent truncation,
Annals of the Institute of Statistical Mathematics, DOI: 10.1007/s10463-015-0526-9

Takeshi Emura
Inst. Stat. Science, Academia Sinica

Joint work with Weijing Wang
Inst. Stat., National Chiao Tung U.

Outlines

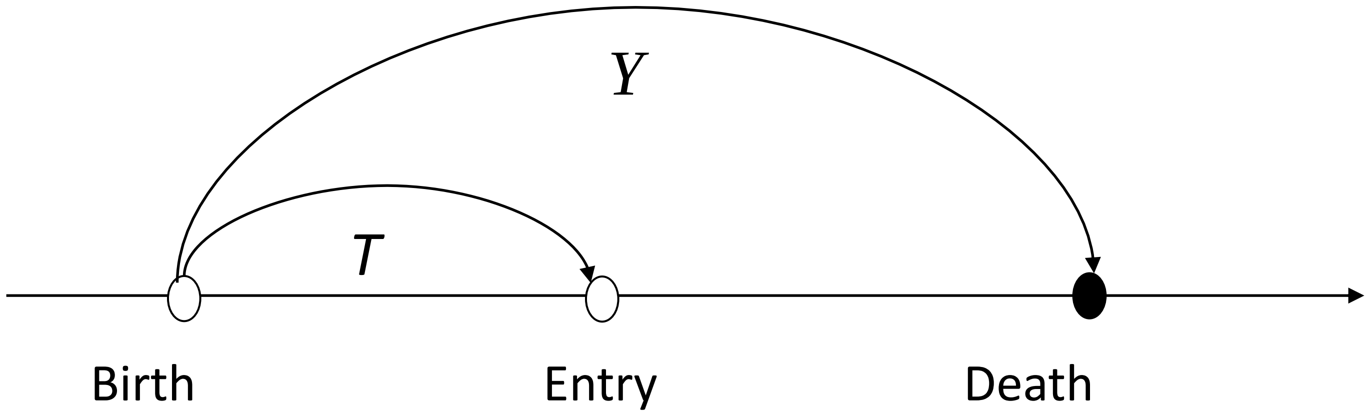
Part I: Review

- Truncated data - i.i.d. case -
- Truncated data - with covariate -
- Existing method - AFT model -

Part II: Proposed method

- Proposed method
- Estimation procedure
- Simulation and data analysis
- Conclusion

Truncation data (i.i.d. case)



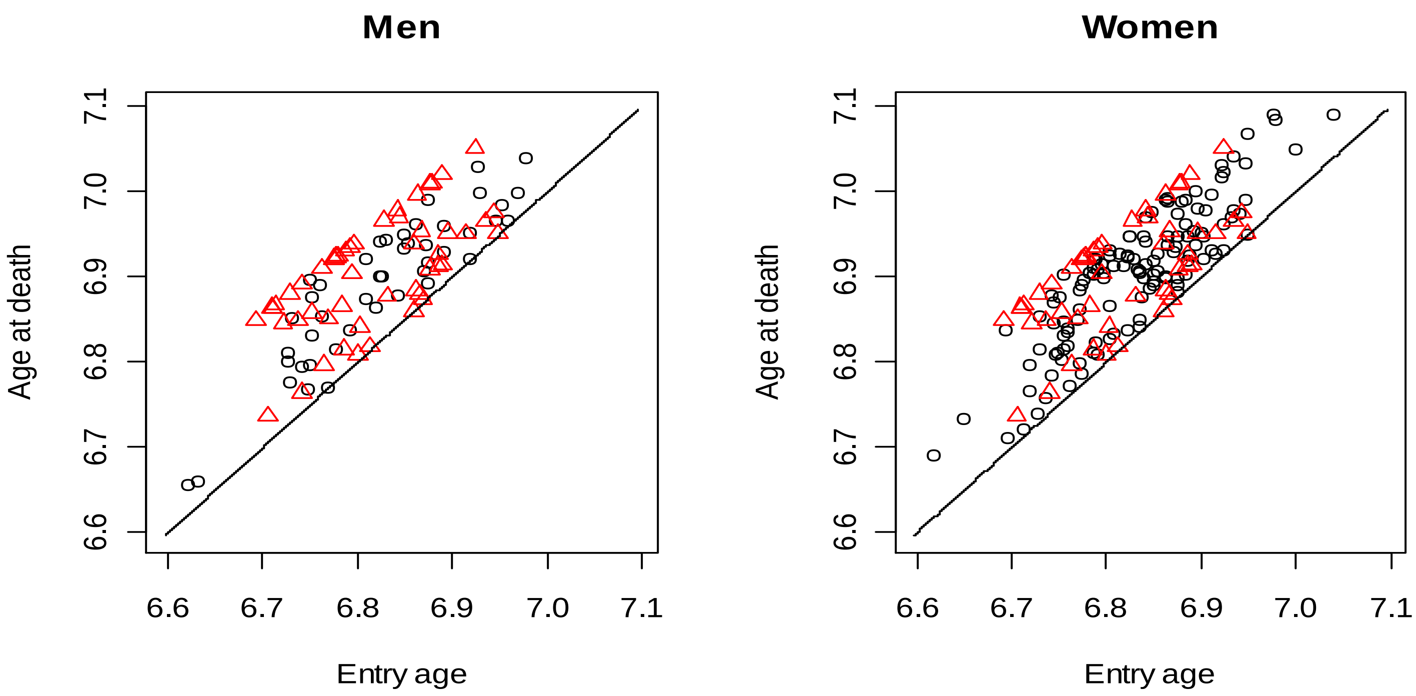
Channing House data (Hyde, 1980)

Available information for Individuals (n=462)

- Entry age
- Age at death or censoring (withdraw)
- Sex (97 man ; 365 women)

Truncation criterion: $T \leq Y$

Truncation data (i.i.d. case)



Δ: Censored individual
O: Died individual

- Hyde (1980) assumed:
  knowing the person's entry age will provide no additional information about prospects for survival
- Under $T \perp Y$, he obtained gender specific survival for Y
  (*T*: Age at entry; *Y*: Age at death)

Truncation data (i.i.d. case)

- Left-truncated data (no censoring, no covariates):

$\{(T_j, Y_j);\ j = 1, \dots, n\}$ subject to $T_j \le Y_j$

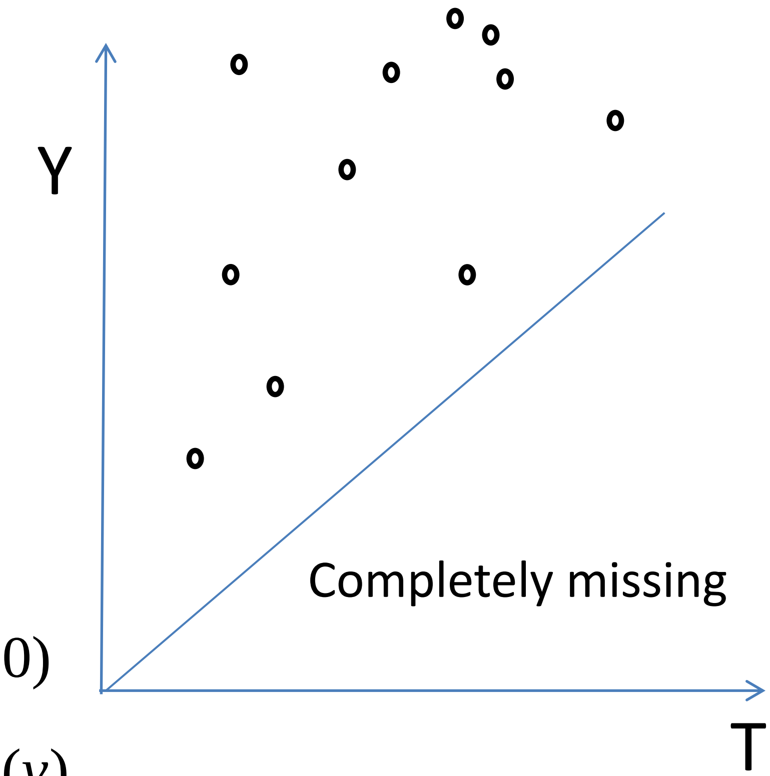
$$\Downarrow$$

i.i.d. from the conditional c.d.f.

$\Pr(T \le x, Y \le y \mid T \le Y)$,

where (T, Y) is

the population random variable



- Quasi-independence assumption (Tsai, 1990)

$$H_0:\ \Pr(T \le t, Y \le y \mid T \le Y) \propto \iint_{\substack{u \le t, v \le y \\ u \le v}} dF_T(u)\, dF_Y(v)$$

Truncation data (i.i.d. case)

Testing quasi-independence assumption

$$\mathrm{H}_0: \ \Pr(T \le t, Y \le y \mid T \le Y) \propto \iint_{\substack{\mathrm{u} \le t, \mathrm{v} \le \mathrm{y} \\ \mathrm{u} \le \mathrm{v}}} dF_T(u) dF_Y(v)$$

Available test statistics:

1. Chen et al. (1996) - Based on Peason-correlation

2. Tsai (1990); Martin & Betensky (2005) - Based on Kendall's tau

$$\mathbf{U_C} = \sum_{i<j} \mathrm{sgn}\{(Y_i - Y_j)(T_i - T_j)\} I(\Omega_{ij}), \quad \text{where } T_i \vee T_j \le Y_i \wedge Y_j$$

- $E[\mathbf{U_C}] = 0 \quad \text{under} \quad H_0$

4. Emura & Wang (2010) - Based on 2 by 2 table

(Optimality under copula-based alternative hypothesis)

NOTE: Reject H_0 → Reject $T \perp Y$

Truncation data with covariates

Left-truncated and right-censored data:

- Y^* : Log-survival time
- T : Log-truncation time
- C : Log-censoring time
- $\mathbf{X}$: *p*-dimensional covariate

Left-truncation:

A pair $(T, Y^*, \Delta, \mathbf{X})$ is observed only when $T \le Y^*$, , where $Y = Y^* \wedge C, \quad \Delta = I(Y^* \le C)$

*If $T = -\infty$, this is usual right-censored data with covariates

➔ fit Cox regression (1972) or AFT regression(Tsiatis, 1990)

Truncation data with covariates

Observed data:

$$\{(T_i, Y_i, \Delta_i, \mathbf{X}_i); (i = 1, \ldots, n)\} \text{ subject to } T_i \le Y_i$$

Rank regression (Lai and Ying, 1991 AS)

Model: Accelerated failure time (AFT) model:

$$Y^* = \boldsymbol{\beta}_0'\mathbf{X} + \varepsilon, \quad \text{where p.d.f of } \varepsilon \text{ is unspecified}$$

Estimation: Log-rank type estimating equation:

$$\mathbf{U}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Delta_i \phi_i(\boldsymbol{\beta}) \left\{ \mathbf{X}_i - \frac{1}{R_i(\boldsymbol{\beta})} \sum_j \mathbf{X}_j I(e_j^T(\boldsymbol{\beta}) \le e_i^Y(\boldsymbol{\beta}) \le e_j^Y(\boldsymbol{\beta})) \right\},$$

where $e_i^T(\boldsymbol{\beta}) = T_i - \boldsymbol{\beta}'\mathbf{X}_i$, $e_i^Y(\boldsymbol{\beta}) = Y_i - \boldsymbol{\beta}'\mathbf{X}_i$,

$$R_i(\boldsymbol{\beta}) = \sum_j I(e_j^T(\boldsymbol{\beta}) \le e_i^Y(\boldsymbol{\beta}) \le e_j^Y(\boldsymbol{\beta}))$$

Truncation data with covariates

Assumptions for Lai & Ying method:

$$\begin{cases} Y^* = \boldsymbol{\beta}_0'\mathbf{X} + \varepsilon \\ (T, C, \mathbf{X}) \perp \varepsilon \quad \cdots(\mathrm{A}) \end{cases}$$

Independent truncation

Why (A) is independent truncation?

By (A), $Y^* - \boldsymbol{\beta}_0'\mathbf{X} \perp T.$

After adjusting the effect of **X**, the truncation variable T contains no information on Y^*

Motivating Example: This model satisfy (A) only under $\rho = 0$

$$\begin{bmatrix} Y^* \\ T \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\beta}_0'\mathbf{X} \\ -1 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right), \quad C \sim N(1, 1)$$

Part II: Proposed method

Proposed method

Relaxing the independent truncation:

Proposed model (dependent truncation liner model):

$$\begin{cases} Y^* = \boldsymbol{\beta}_0'\mathbf{X} + \gamma_0 T + \varepsilon \\ \quad (T, C, \mathbf{X}) \perp \varepsilon \end{cases} \quad \ldots. \qquad \text{(B)}$$

NOTE: Special case of $\gamma_0 = 0$, ➔ Lai & Ying model

Example: Bivariate normal model

$$\begin{bmatrix} Y^* \\ T \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\beta}_0'\mathbf{X} \\ -1 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right), \quad C \sim N(1, 1)$$

This model satisfy (B) with

$$Y^* = \boldsymbol{\beta}_0'\mathbf{X} + \gamma_0 T + \varepsilon, \quad \text{where } \rho = \gamma_0 \quad \text{and} \quad \varepsilon \sim N(\gamma_0, 1 - \gamma_0{}^2)$$

Estimation procedure

Setting:

- Model : $\begin{cases} Y^* = \boldsymbol{\beta}_0'\mathbf{X} + \gamma_0 T + \varepsilon \\ (T, C, \mathbf{X}) \perp \varepsilon \end{cases}$ .... (B)
- Left-truncated & right-censored data :

$$\{(T_i, Y_i, \Delta_i, \mathbf{X}_i); (i = 1, ..., n)\} \text{ subject to } T_i \le Y_i$$

Interest:

1) Joint estimation of $(\boldsymbol{\beta}_0', \gamma_0)$

2) Estimation of $S_\varepsilon(t) = \Pr(\varepsilon > t)$

Estimating equations for

a) $\boldsymbol{\beta}_0$ ➔ Inverting the log-rank test statistics

b) γ_0 ➔ Inverting the quasi-independence test statistics

Estimation procedure

Residual transformation:

$$\begin{cases} \varepsilon_i^Y(\boldsymbol{\beta},\gamma) = Y_i - \boldsymbol{\beta}'\mathbf{X}_i - \gamma T_i \\ \varepsilon_i^T(\boldsymbol{\beta},\gamma) = T_i - \boldsymbol{\beta}'\mathbf{X}_i - \gamma T_i \end{cases}$$

a) Log-rank estimating equation:

By assumption (B), $H_0: Y^* - \boldsymbol{\beta}_0'\mathbf{X} - \gamma_0 T \perp \mathbf{X}$ is true.

$$\mathbf{S}_n^{Logrank}(\boldsymbol{\beta},\gamma)$$

$$= -\sum_{i<j}(\mathbf{X}_i - \mathbf{X}_j)\,\mathrm{sgn}\{(\varepsilon_i^Y(\boldsymbol{\beta},\gamma) - \varepsilon_j^Y(\boldsymbol{\beta},\gamma)\} I\{\breve{\varepsilon}_{ij}^T(\boldsymbol{\beta},\gamma) \le \tilde{\varepsilon}_{ij}^Y(\boldsymbol{\beta},\gamma)\} O_{ij}(\boldsymbol{\beta},\gamma)$$

b) Quasi-independence estimating equation:

By assumption (B), $H_0: Y^* - \boldsymbol{\beta}_0'\mathbf{X} - \gamma_0 T \perp T - \boldsymbol{\beta}_0'\mathbf{X} - \gamma_0 T$ is true

$$S_n^{Kendall}(\boldsymbol{\beta},\ \gamma)$$

$$= \sum_{i<j}\mathrm{sgn}\{(\varepsilon_i^T(\boldsymbol{\beta},\ \gamma) - \varepsilon_j^T(\boldsymbol{\beta},\ \gamma))(\varepsilon_i^Y(\boldsymbol{\beta},\ \gamma) - \varepsilon_j^Y(\boldsymbol{\beta},\ \gamma))\} I\{\breve{\varepsilon}_{ij}^T(\boldsymbol{\beta},\ \gamma) \le \tilde{\varepsilon}_{ij}^Y(\boldsymbol{\beta},\ \gamma)\} O_{ij}(\boldsymbol{\beta},\ \gamma).$$

Martin & Betensky type statistic (2005)

Estimation procedure

Regression estimator :

$$\begin{pmatrix}\hat{\boldsymbol\beta}\\ \hat{\gamma}\end{pmatrix}: \quad \begin{cases}\mathbf{0}=\mathbf{S}_n^{Logrank}(\boldsymbol\beta,\gamma)\\ 0=S_n^{Kendall}(\boldsymbol\beta,\gamma)\end{cases}$$

⇦ Non-monotonic step functions

Use "sequential grid search" to find a minimum of

$$M_n(\boldsymbol\beta,\gamma)=\|\mathbf{S}_n^{Logrank}(\boldsymbol\beta,\gamma)\|_1+|S_n^{Kendall}(\boldsymbol\beta,\gamma)|$$

NOTE: Newton-Raphson, bisection method, and linear programing (Jin, Lin and Wei, 2003) do not work: Brown & Wang (2005) ??

Error distribution $S_\varepsilon(t)=\Pr(\varepsilon>t)$ **:**

$$\hat{S}_\varepsilon(t;\hat{\boldsymbol\beta},\hat{\gamma})=\prod_{u\le t}\left\{1-\frac{\sum_j I(\varepsilon_j^Y(\hat{\boldsymbol\beta},\hat{\gamma})=u,\Delta_i=1)}{\sum_j I(\varepsilon_j^T(\hat{\boldsymbol\beta},\hat{\gamma})\le u\le\varepsilon_j^Y(\hat{\boldsymbol\beta},\hat{\gamma}))}\right\}$$

⇨ $\Pr(\exp(Y^*)>t\mid\mathbf{X},T)=\hat{S}_\varepsilon(\log(t)-\hat{\boldsymbol\beta}'\mathbf{X}-\hat{\gamma}T)$

Subject-specific
Survival curve

Estimation procedure

Theorem 2 (manuscript) :

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0, \hat{\gamma} - \gamma_0) \quad \rightarrow \quad N(\mathbf{0}, \mathbf{A}_0{}^{-1}\mathbf{B}_0\mathbf{A}_0{}^{-1})$$

Empirical variance estimator :

$$\hat{\mathbf{A}}_0 =$$

$$\iint \mathbf{D}[h\{(t_1, y_1, \delta_1, \mathbf{x}_1), (t_2, y_2, \delta_2, \mathbf{x}_2); \hat{\boldsymbol{\beta}}, \hat{\gamma}\}] dF_n(t_1, y_1, \delta_1, \mathbf{x}_1) dF_n(t_2, y_2, \delta_2, \mathbf{x}_2)$$

(Kernel estimator: Uniform kernel with bandwidth : $b = n^{-1/3}$)

$$\hat{\mathbf{B}}_0 = \sum_{i=1}^{n} \phi_{F_n}(F_{(j)} - F_n; \hat{\boldsymbol{\beta}}, \hat{\gamma}) \phi_{F_n}(F_{(j)} - F_n; \hat{\boldsymbol{\beta}}, \hat{\gamma})' / n$$

Theorem 3 (manuscript) :

$$\sup_{t \in [a,b]} | \hat{S}_\varepsilon(t; \hat{\boldsymbol{\beta}}, \hat{\gamma}) - S_\varepsilon(t) | \xrightarrow{P} 0, \quad \text{where} \quad S_\varepsilon(t) = \Pr(\varepsilon > t)$$

Simulation

Model :

$$\begin{bmatrix} Y^* \\ T \end{bmatrix} \sim N\left(\begin{bmatrix} \beta_0 X \\ -1 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), \quad C \sim N(1,1), \quad X \sim U(0,1)$$

This induces the linear regression model

$$Y^* = \beta_0 X + \gamma_0 T + \varepsilon, \quad \text{where } \rho = \gamma_0 \text{ and } \varepsilon \sim N(\gamma_0, 1-\gamma_0{}^2)$$

Parameter configurations:

	1	2	3	4	5	6
(β_0, γ_0)	(0, -0.5)	(0, 0)	(0, 0.5)	(1, -0.5)	(1, 0)	(1, 0.5)
$\Pr(T \le Y^*)$	0.72	0.76	0.84	0.80	0.85	0.84
$\Pr(C < Y^* \mid T \le Y^*)$	0.30	0.27	0.23	0.41	0.39	0.34

Table 1. Simulation results for the proposed estimator $(\hat{\beta}, \hat{\gamma})$ †

(β_0, γ_0)	n	Bias	SD	SDE	95%Cov
(0, -0.5)	150	(-0.015, -0.002)	(0.315, 0.167)	(0.335, 0.166)	(0.960, 0.950)
	300	(-0.003, -0.006)	(0.227, 0.117)	(0.230, 0.113)	(0.960, 0.955)
(0, 0)	150	(0.001, -0.023)	(0.394, 0.164)	(0.402, 0.176)	(0.955, 0.970)
	300	(0.013, -0.008)	(0.305, 0.105)	(0.285, 0.119)	(0.935, 0.970)
(0, 0.5)	150	(-0.004, -0.014)	(0.363, 0.126)	(0.347, 0.131)	(0.920, 0.955)
	300	(-0.010, -0.006)	(0.230, 0.090)	(0.239, 0.087)	(0.955, 0.935)
(1, -0.5)	150	(0.014, -0.014)	(0.379, 0.160)	(0.349, 0.159)	(0.920, 0.935)
	300	(0.006, -0.012)	(0.257, 0.110)	(0.242, 0.110)	(0.920, 0.950)
(1, 0)	150	(0.001, 0.003)	(0.429, 0.158)	(0.408, 0.160)	(0.915, 0.965)
	300	(0.006, -0.011)	(0.297, 0.108)	(0.287, 0.110)	(0.940, 0.975)
(1, 0.5)	150	(0.052, -0.001)	(0.353, 0.108)	(0.346, 0.124)	(0.935, 0.980)
	300	(0.025, 0.002)	(0.234, 0.070)	(0.238, 0.080)	(0.960, 0.970)

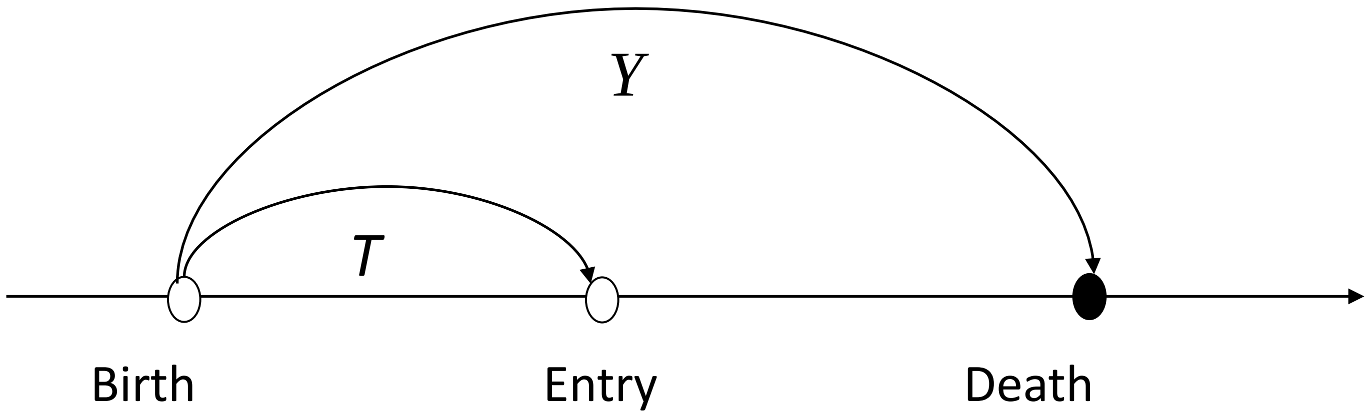
† The average of biases (Bias), standard deviation (SD), the average of standard error estimate (SDE), and the empirical coverage probability of 95% confidence interval (95%Cov) based on 200 simulation runs are reported.

Table 2. Simulation results for the proposed estimator $\hat{S}_{\varepsilon}(t;\hat{\beta},\hat{\gamma})$ †

(β_0,γ_0)	n	Bias $(S_{\varepsilon}(t_{0.25}), S_{\varepsilon}(t_{0.50}), S_{\varepsilon}(t_{0.75}))$	Standard deviation $(S_{\varepsilon}(t_{0.25}), S_{\varepsilon}(t_{0.50}), S_{\varepsilon}(t_{0.75}))$
(0, -0.5)	150	(-0.066, 0.008, 0.068)	(0.089, 0.143, 0.113)
	300	(-0.075, 0.000, 0.071)	(0.062, 0.103, 0.082)
(0, 0)	150	(-0.001, -0.009, -0.021)	(0.105, 0.141, 0.133)
	300	(-0.002, -0.005, -0.012)	(0.073, 0.102, 0.090)
(0, 0.5)	150	(0.044, 0.000, -0.043)	(0.103, 0.134, 0.139)
	300	(0.041, -0.001, -0.044)	(0.071, 0.096, 0.106)
(1, -0.5)	150	(-0.074, -0.008, 0.057)	(0.095, 0.154, 0.119)
	300	(-0.079, -0.007, 0.066)	(0.066, 0.109, 0.081)
(1, 0)	150	(0.011, 0.009, -0.005)	(0.105, 0.134, 0.128)
	300	(0.000, -0.005, -0.009)	(0.073, 0.096, 0.088)
(1, 0.5)	150	(0.038, -0.009, -0.055)	(0.092, 0.110, 0.111)
	300	(0.041, -0.002, -0.049)	(0.072, 0.085, 0.082)

† The average of biases (Bias) and standard deviation based on 200 simulation runs are reported. The true values of $(S_{\varepsilon}(t_{0.25}), S_{\varepsilon}(t_{0.50}), S_{\varepsilon}(t_{0.75}))$ are (0.75, 0.5, 0.25) respectively.

Data analysis



Channing House data (Hyde, 1980)

Available information for Individuals (n=462)

- *T* : Entry age
- *Y** : Age at death or censoring
- *X*: Sex (97 man ; 365 women)

Linear model:

$$Y^* = \beta_0 X + \gamma_0 T + \varepsilon, \quad \text{where } \varepsilon \text{ is unspecified}$$

Data analysis

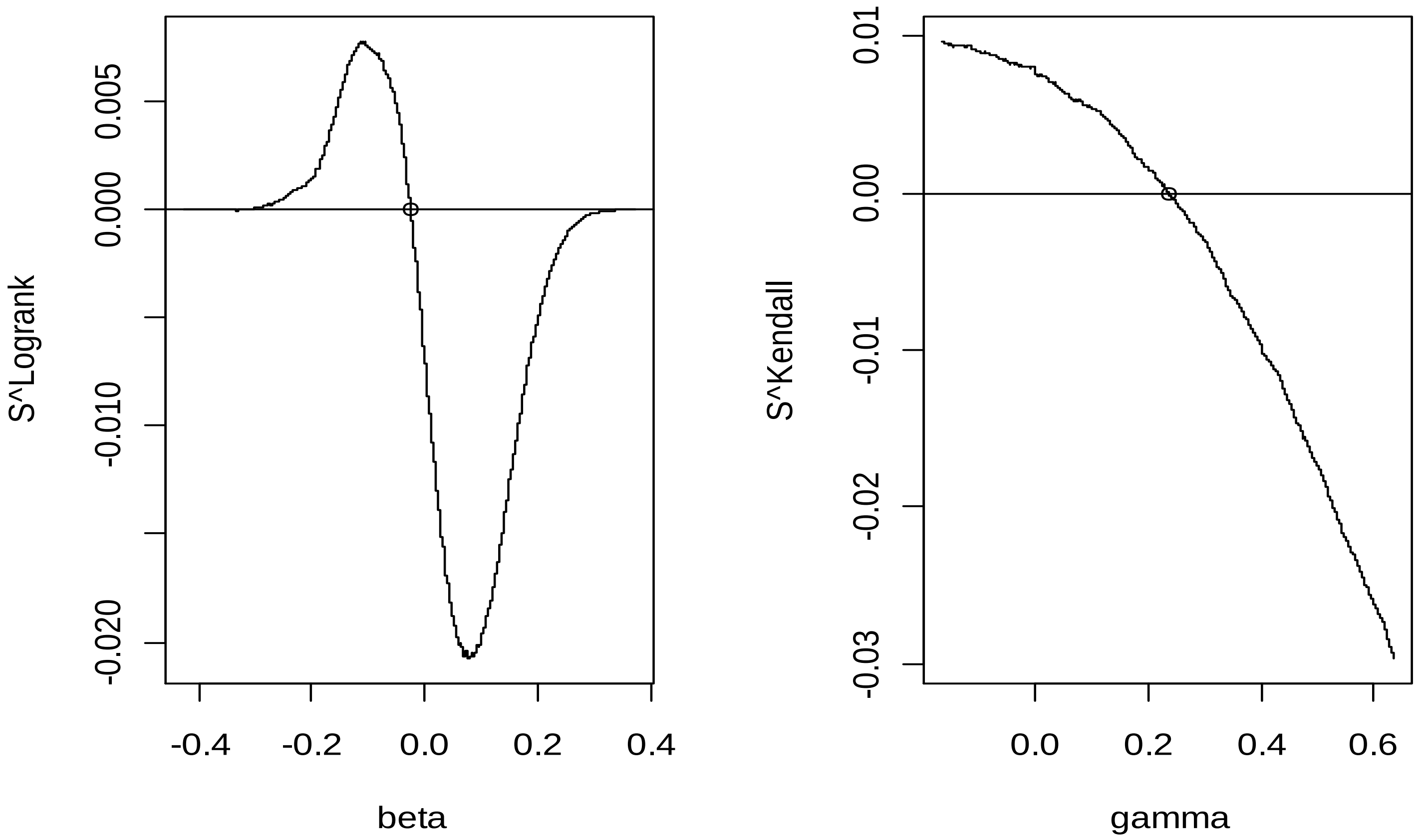


Fig. 3. Plots of $S_n^{Logrank}(\beta, \hat{\gamma})$ and $S_n^{Kendall}(\hat{\beta}, \gamma)$ based on the Channing house data. The numerical solutions $\hat{\beta} = \text{-}0.026$ and $\hat{\gamma} = 0.236$ obtained from the grid search algorithm are indicated by "o".

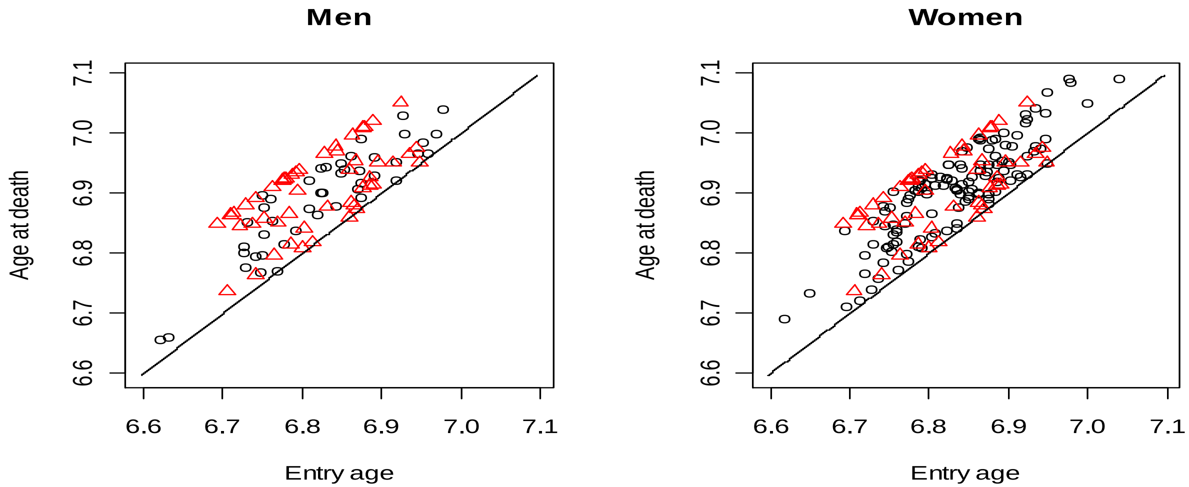
Data analysis

Interpretation:

$$\text{Age at death} = -0.026\times\text{Gender} + 0.236\times\text{Age at entry} + \text{Error}$$

$$\begin{cases} -0.026 & \text{....95\% conf. interval (-0.115, 0.063)} \\ 0.236 & \text{....95\% conf. interval (0.010, 0.461)} \end{cases}$$

Late entry to Channing house prolong the survival



Δ: Censored individual

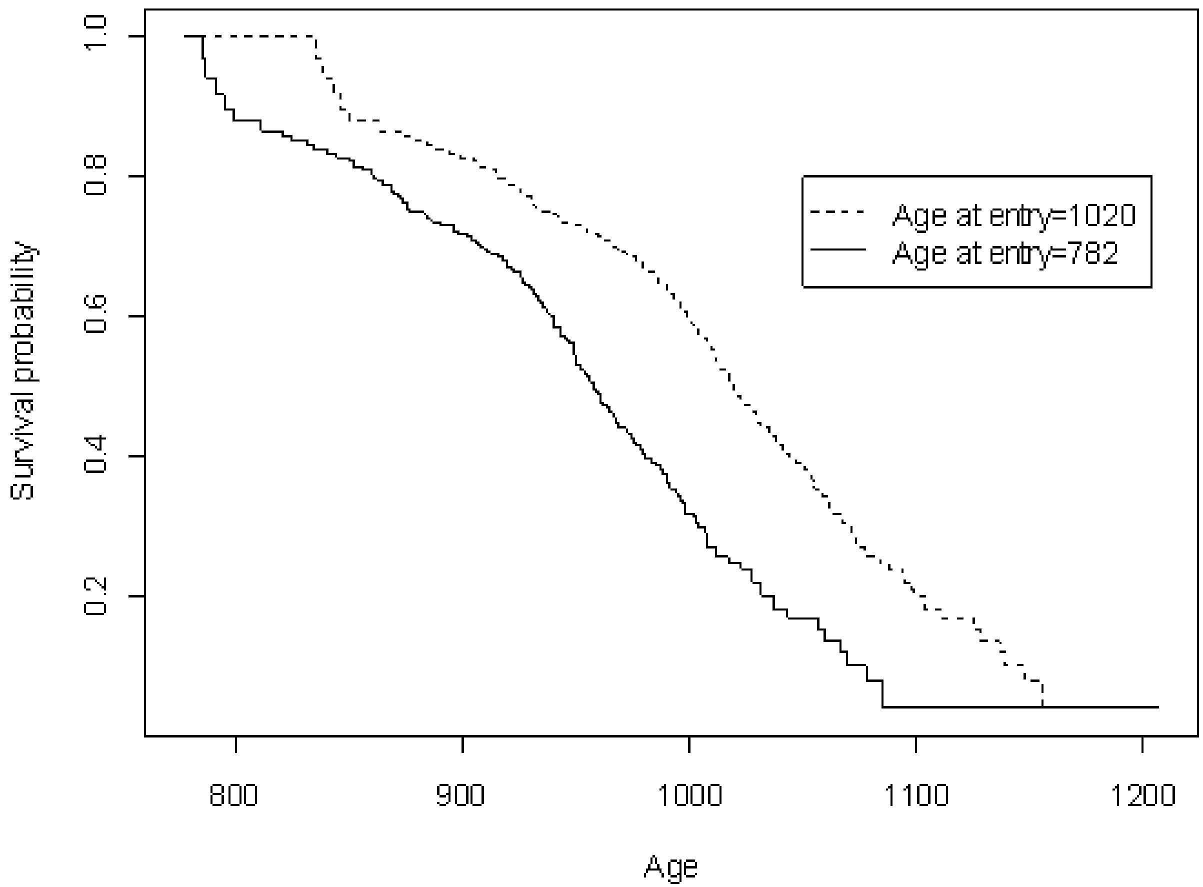
O: Died individual

Data analysis

Subject specific survival :

Predicted survival for the two individual:

- ID#1: Entry age = 782 (month), sex = male
- ID#2: Entry age =1020 (month), sex = male



Conclusion

- We propose a semi-parametric AFT model which utilizes *both* covariates and truncation variable to model the survival time.

- The model is an extension of Lai & Ying (1991) model that can only utilize covariate as regressors.

- In Channing house data, the entry age (truncation variable) is shown to be informative in the survival prediction.

Thank you for your kind attention